REPORT OF THE 2nd AQUAINT DIALOGUE EXPERIMENT

Jean Scholtz and Emile Morse
National Institute of Standards and Technology
jean.scholtz@nist.gov; emile.morse@nist.gov
June, 2003

Executive Summary

The second dialogue experiment showed considerable progress in both system development and experimental methods. Longer and more realistic scenarios provided the opportunity to see more in-depth use of the systems. Improved systems and more experienced participants allowed us to move from syntactic analysis to usage analysis.

Four systems participated. Two systems had web-based user interfaces that provided additional features to the users. Although we retained the use of wizard and system modified responses, systems were able to handle a higher percentage of responses either alone or as system-modified responses than in experiment one. The use of a common set of documents also provided a better experimental comparison.

We used Camtasia to record the screens of the participants and were able to playback these screens after the sessions for a retrospective discussion of the participant's behavior.

Observations of the additional features provided by the web-based user interfaces have caused us to redefine "dialogue." For future experiments we suggest that dialogue be defined as: requests to the system for content + exploration of content. Exploration of content can be considered as "local interactions." That is, systems may provide the user with visualizations or other types of views, or interactions with existing (already located) content. Requests for content can be made explicitly by the user; can be facilitated by the system; or can be delivered automatically by the system. Metrics for the next experiment should take these factors into account.

Introduction

The goal of this second Aquaint Dialogue experiment was to continue to investigate the types of dialogue that analysts would expect to engage in with an information system. We were also particularly interested in refining the metrics used to assess the systems. This experiment differs from the first study in several ways, based on the suggestions of the participants in the first study.

- Longer scenarios: In the first study, 10 scenarios were presented to the subjects and each was worked on for 15 minutes. To address this problem, more complicated scenarios were developed that could be worked on for at least one hour. Two scenarios were presented to each subject.
- More realistic data and scenarios: In the first study, participants were asked to collect relevant data for each scenario. In the current study, the CNS collection was used. In addition, scenarios were developed by CNS analysts.
- Subjects with more domain expertise: The first study employed TREC assessors who had analytic experience. For the current study, we recruited some Navy reservists who performed analysis as their duty; we supplemented these with TREC assessors.
- Web-based interfaces: In the first study, NIST developed a chat interface that was used for all the interactions between the test subject and the systems. In this iteration, it was decided that alternative native system interfaces could be used as long as they were web-based.
- Better metrics: We have switched from using 7-point Likert scales to semantic scales and have added more quantitative measures.

Methodology

There were four systems participating in this pilot evaluation. There were also four subjects. Six scenarios were chosen from the set provided by CNS; two short scenarios were chosen for Training. Each scenario as delivered by CNS had a partial decomposition that included the subset of data relevant to the scenario. We used this information to create two sets of tasks; each task set had one task that was more focused and could be answered by referring to a single sub-collection and the other task relied on more than one sub-collection for arriving at the best outcome. The order of the tasks within the sets was constant. All scenarios were given to the researchers about one month before the scheduled experiments. Appendix A contains the scenarios for both question sets including the Training scenarios.

Each experiment began with a short period of about 10 minutes during which subjects were given a description of their overall task (Appendix B) and printed instructions for using the interface. The interface descriptions were tailored for the particular web-based or NIST chat that was employed during the test. There was an observer/test administrator present during the entire time whose role was to assist during training and to make notes for later debriefing. Analysts logged into the system and were given the Training scenario to read. Thirty minutes were devoted to training with the system during which the subject submitted queries and learned to interpret the system's responses. They also got used to saving relevant information in Notepad, Word, or on paper. They learned the use of the Success/Failure buttons that were used to terminate the session and they learned to use the Questionnaire (see Appendix C) that was to be used at the end of each scenario. This process from logging into the system to answering the questionnaire was repeated for the two experimental scenarios except that the duration for each was one hour. After all the experimental scenarios were complete, the observer reviewed her notes with the subject to gain additional insight into the analytical process and to clarify any questions. A playback facility was used to view the subject's screen during the review process. The debriefings were audio-taped and later transcribed.

For each scenario (both training and experimental), log files were created using the standard format (http://www.itl.nist.gov/iaui/894.02/projects/aquaint/dialogue/aquaint-file-spec.html) developed and used in the first pilot study. When the NIST chat interface was used, the logs were created at NIST; when other web-based interfaces were used, the logs were created at that location and were sent to NIST immediately after the experiment was complete. In addition, Camtasia (http://www.techsmith.com) was used to capture the screen of the analyst's computer during the entire session.

Each research team was given a copy of the observer's notes supplemented with the audio transcription of the debriefing; a copy of the log files (for those using the NIST chat interface only); and a spreadsheet with the quantitative and subjective rating results for their system. This report contains an overall summarization of the results from the experiment.

System Descriptions

Two of the systems used the NIST text chat tool developed for experiment 1 (referred to as system 1 and system 3 in the remainder of this report). The other two systems supplied their own web-based user interface but retained the logging format developed by NIST for the first experiment.

System 2 provided a multi window user interface. Components of the user interface were a User Query Tab, a Dialogue Tab, an Answer Tab, a status bar, and a Visual tab. The User Query tab is used to enter original queries. The Dialogue tab is used to converse with the system if rephrasing is necessary. The Answer tab is where the answer to the query appears and follow-up actions can be initiated. The Status Bar is a progress indicator. TheVisual tab displays a visual representation of the query and the best answers for that query.

System 4 also provided multiple panes in the user interface. In addition to a Text input pane, a Conversation pane, and a Status pane, a 'Related Question' list was provided. After the user posed a question, similar questions were displayed here. The user could "ask" one of these questions by merely clicking on it.

Results

What did we learn about the systems

Quantitative results

Each system was used by two analysts and each analyst performed two scenarios. Preliminary observations showed that there were no significant differences between the pair of subjects or between the pair of scenarios; standard error values were approximately 10% of the mean. The four values were averaged to produce the results shown in Table 1.

Table 1: Quantitative Measurements (/scenario)

	System 1	System 2	System 3	System 4
Total responses	36.5	39.0	32.5	92.5
# for the analyst	15.0	12.3	16.5	29.3
# for the system	21.5	26.8	16.0	63.3
system responses	1.0	26.8	8.0	54.5
wizard responses	0.3	0.0	5.5	2.3
modified responses	20.3	0.0	2.5	6.5
Turns	9.5	10.8	13.8	27.3
Threads	2.8	1.5	2.3	1.0
Wait time (min)	13.2	10.8	30.5	4.4
Time (min)	56.4	47.4	57.0	48.8
Success	4/4	4/4	4/4	2/4

System 4 had significantly more total responses -- both the analyst and the system engaged in more exchanges. System 1 relied heavily on modified responses, while the other systems used this sparingly. System 3 used wizard responses significantly more than the other systems.

The total number of responses during a single scenario averaged between 32 and 92 indicating that exchanges were happening at a high rate. Analysts' responses accounted for less than half of the total responses. The system responses by systems 1 and 3 were lower than either of the other systems and system 2 had significantly fewer than system 4.

Of the system responses, there is a clear difference in the type of these responses as captured in the log files. System 1 produced almost only 'modified' responses, while the most common type of response for the other systems was 'system'.

When compared with the results of the first pilot study, it seems that the systems are relying much less on 'wizard' responses. We attribute this to the maturation of the systems.

A 'turn' was defined as an analyst response followed by a system response. These values appear to roughly track with the number of analyst responses, which is to be expected. 'Threads' are defined as the subject taking a new direction or using a new line of reasoning. In one of the systems (#2), this marker was implicit and relied on the subject entering a new query; the other three systems a user-selectable option which provided an explicit marker.   The low number of threads indicates that analysts were able to follow a particular line of questioning with the systems.

In the first pilot, when scenarios were only 15 minutes long, subjects complained about the systems taking too long; indeed, subjects spent half the time just waiting for a response. In the current experiment, the systems had a much lower average wait time.

Subjects believed that their task was successful in 14 out of 16 scenarios. During the discussions that took place after the first study, we found that both the analysts and the researchers were not happy with the binary value of this variable. The analysts wanted to be able to express something less than complete success and/or something shy of total failure. The researchers indicated a preference for a confidence scale to capture the same type of distinctions. Confidence rating scales have been added to the questionnaire and are presented in the following section.

Subjective ratings

Nine of the ratings asked the analysts to assess the dialogue that they carried out with the system. Seven-point semantic scales were used. For eight of these scales, the 'best' score was 1 and the 'worst' was 7. The mean values for these factors are shown in Table 2. In general, systems 1 and 3 had lower, 'better' scores for each factor than systems 2 and 4. In fact, the differences for many of the factors reached statistical significance. It is pertinent to note that systems 1 and 3 used the NIST Chat Interface, while systems 2 and 4 used web-based interfaces developed specifically for them. The findings imply that the simpler NIST Chat interface was easier to use that the relatively more complex custom interfaces. As systems begin to develop complex, custom interfaces the role of training will become more important.

Table 2: Subjective Ratings for Dialogue Factors

1	**System 1**	**System 2**	**System 3**	**System 4**	7
Pleasing	2.0	3.8	1.8	3.3	Irritating
Helpful	2.0	3.5	1.8	2.8	Hindrance
Understood my questions	1.3	3.3	3.0	3.3	Did not understand
Easy to understand	2.0	3.5	2.0	2.5	Incomprehensible
Good context	1.8	3.8	2.0	3.0	Bad context
Natural dialogue	3.0	3.8	1.8	2.8	Unnatural
Knew what to say	1.8	3.3	1.8	3.5	Hard to know what to say
Natural turn-taking	2.3	4.0	1.5	2.0	Hard to know when to speak

Table 3 shows the mean values of satisfaction questionnaire ratings where 'best' is not necessarily associated with either end of the scale but rather is somewhere in the middle. Verbose conversation is not any better than terse conversation. Verbosity/terseness is the ninth of the dialogue factors; it was not included in the previous table because of the difference in interpretation of the scale. System 3 was significantly more verbose than the other systems.

The locus of control – the system or the analyst – is similar to the verbosity scale in that the 'ideal' is likely to be somewhere in the middle rather than at an endpoint. There was no difference between systems on this variable; subjects, in general, felt that they were in control.

Table 3: Elements that have a neutral balance point

1	**System 1**	**System 2**	**System 3**	**System 4**	7
Verbose	3.8	4.5	2.5	3.5	Terse
System in control	5.5	5.0	6.5	5.5	Self in control

The analysts rated their prior level of knowledge with respect to the particular scenario by answering the questions shown in Table 4. It is interesting to note that the naval reservists did not rate their expertise higher that the retired analysts. There were 2 instances among the scenarios (including training) where an analyst rated his knowledge 'Expert'. These scenarios just happened to be the area in which they had/have their expertise. They consistently reported that they could apply little expertise since they did not possess much.

Table 4: Expertise Factors

1	**System 1**	**System 2**	**System 3**	**System 4**	7
Expert	4.8	5.5	5.8	3.3	Novice
Applied much knowledge	5.3	5.0	5.0	6.3	Little

The Confidence rating scales were added, as mentioned in the Introduction, to try to overcome the black-and-white nature of the success/failure decision. By and large, the analysts were very confident in both the accuracy and coverage of the information. There was no difference among the systems.

Table 5: Confidence ratings

1	**System 1**	**System 2**	**System 3**	**System 4**	7
Not confident in accuracy	5.8	5.3	5.8	5.8	Very confident
Not confident in coverage	6.3	5.3	5.8	4.5	Very confident

Whether an analyst would use the system is an indicator of his/her overall satisfaction with the system. This question was posed twice in the questionnaire but the mapping of 'best' and 'worst' was reversed. Although the duplication was accidental, it provided an opportunity to assess the sensitivity of the measure. When the question was posed with 'would use' as 7, all of the systems had values above the midpoint; conversely, when 'would use' mapped to 1, all of the values were below the midpoint. This indicates that the subjects were consistent in these ratings. There were no differences between any of the systems when compared statistically.

Table 6: Overall Satisfaction

1	**System 1**	**System 2**	**System 3**	**System 4**	7
Would NOT use	4.8	5.5	6.0	4.5	Would definitely use
Would use regularly	3.0	3.0	1.5	3.5	Would never use

What the analysts wrote about the systems

For each subjective rating element, the subjects were allowed and encouraged to make comments. This section presents those comments. Subjects did not comment about 'Expertise' or 'Confidence' factors.

Comments about dialog factors (include terse/verbose)

There were no comments about two of the factors – 'Easy to understand/Incomprehensible' and 'Natural turn-taking/hard to know when it was my turn'.

Pleasing/Irritating

- "a little slow – another interface would be good to pursue more than one track at a time. A bell or some such thing to indicate when a response is received would be nice."
- "As I get used to the system, I like it more and more."
- "The more I get used to the system, the better I like it. There are a number of paths in the system that take time to understand and use."

- “Somewhat confusing with responses to different systems popping up. However, responses include original question and which system – which is good.”
- “I like how the question being answered is restated. Would be better still if the beginning of each set of responses were highlighted.”

Helpful/Hindrance

- “Yes, but it did not get far enough into the test to rate it higher.”
- “But when a worded response is included – not just URLs.”

Understood my questions/not

- “’How often…’ and ‘Any other occurrences of …’ yielded different responses, even though the questions were basically the same”
- “Notice with statements it was harder for the system to give the right response I was looking for.”
- “During] This scenario system said it was] unable to understand my question and asked me] to rephrase.”
- “Query was lost.”
- “Two questions were put together as one.”
- “Generally OK. Had to ask for an identification twice, after I already found a prior reference to an individual.”

Good context/not

- “I think the answer is yes, but I did not complete the test.”
- “relevance of answer to scenario.” ?]

Natural dialogue/unnatural

- “The system did miss or lose one query.”
- “I would suggest if the system does not understand the question then ask for clarification.”

Easy to understand what to say/difficult

- “Not sure yet.”
- “Need to highlight references; especially important in longer documents.”

Verbose/terse

- “Lots of material to look through.”

Comments about overall satisfaction (Would use/Would never use)

“Would use it but would need the source of where the data was coming from for validation.”

“Need to highlight references; especially important in longer documents.”

“I would need dates and sources of information given. Assuming the system can give me that, when asked, I would definitely use it.”

“I think, given a week’s time, I could develop enough information to write a report.”

“The information seems worthy for my report but need to know where info came from and its validity.”

“Yes, but again would like to validate the information.”

“Answers contained the info necessary to write response to the scenario question.”

“Yes, but since I did not complete the task, I am not sure how well the system performed.”

“The system should improve the productivity of an intelligence analyst significantly and improve the content of reports.”

"The system was very convenient, easy to use, provided additional questions. Although I considered the task a failure in that it did not answer the question, the system did everything it should have to provide the potential answer."

"There were no 'related questions' for about 80% of the session. I missed them. The session was somewhat a success but not 100%."

"Generally, found excellent source materials. Need to better identify primary sources, as this provides vital context to analysts."

"Tool was unable to locate/display important data for adequate answer to this scenario."

"Most reports require quick turn-around and further steps would inhibit the effectiveness of the report."

"With increased usage I may use this system in the future."

"I like that it came up with lots of answers. I especially like getting URLs/original documents so that I can judge reliability, relevance, etc."

"There were quite a few irrelevant responses (system 1?) and many repetitious."

What we learned about the analysts

Queries

Appendix D contains a table of the initial queries for each scenario and each analyst. Of the 16 initial queries, only one was posed as a sentence fragment/phrase. There were only 3 spelling errors. These findings are at odds with the results of the first Dialogue pilot study. In that study, subjects mistyped, used incomplete sentences, used colloquial phrases, and often made statements rather than asking questions. The scenarios used in this study were more complex than those used in the first study, where we found that subjects would enter the scenario content nearly verbatim as the initial query. The complexity seems to have forced the analysts to break the problems up before issuing any requests for information. In other respects, these initial queries confirm our previous observations; e.g., analysts often provide a context for their query.

Appendix E contains a full listing of the text that the analysts submitted to the system for each scenario. Analysts used relatively well-formed queries throughout their interaction with the systems; grammar and spelling errors were infrequent. Only 23 spelling errors were encountered during the 16 hours of data collection. The most common grammar errors were in terminal punctuation, i.e., use of a period instead of a question mark. There were several instances of subject-verb disagreement. By and large, the analysts provided the systems with correct input and their performance did not decrease significantly as the session proceeded.

Dialogue

What the analysts Said and Did (Observers Logs and Debrief Transcription)

During the time when the subjects were actively working with a system on a scenario, the observer refrained from asking questions. The analysts were not asked to perform in a talk-aloud fashion. Therefore the observer took notes and flagged times when she would have liked to ask a question. During the debriefing, the questions were asked. The playback of the analysts' sessions were used if needed in the debriefing session. The final observer logs were composed of a transcription of the hand-written notes, a set of coded comments based on the debriefing and a set of codings based on overall observations. The following list is a composite of the common findings.

1. Systems should try not to present answers or documents that analysts have already seen.
2. Attributions for material are essential. Date information is also critical.
3. Analysts are 'unstoppable'; they optimize their use of time. Wait time can be effectively used.
4. Analysts test the system by asking questions they already know the answer to. The answer may come from their prior knowledge or the answer may be contained in a document the system has already delivered in response to a different query.

5. Answers or summaries should accurately reflect the documents they stand for. Analysts also wanted some verification or indication in the document of the keywords or concepts that were deemed relevant to the query.
6. Variant spellings and unwarranted equivalences (USSR is not the same as Russia) are issues.
7. When the system rephrases the analyst's question, the analyst should be allowed to confirm whether the modified query is equivalent to the original query.
8. When a system responds to an analyst's query and asks if the analyst would like to see more documents, the number of documents available should be explicitly stated if known.
9. Analysts need to be able to save content information and attribution information returned by the system.
10. Analysts need to know what they have done during a session and would like to be able to save these traces.
11. Analysts need to know when the information they are looking for doesn't exist.
12. Confidence in the system increases or decreases with use. Confidence increases as analysts find "nuggets" of information. Confidence decreases as systems fail the testing done by analysts. Confidence also decreases when analysts find no basis for the return of a document in response to a particular query.

What we learned about the scenarios

The scenarios that were used in this study were created by CNS. CNS also provided partial decompositions. There were five or six areas in each of the decompositions which represent key factors that would be addressed in a final report based on the scenario. As shown in Appendix A, the subjects were shown the suggested decomposition at the end of the questionnaire. They were asked to check 'Yes' if the element was something which they had considered or had been planning to consider during their investigation and 'No' if it was not. Table 7 shows the results. The Question # is a concatenation of the question set (A or B) and the order in which the question was presented.

The analysts agreed with the CNS decompositions for both questions in set B and with the second question in set A. However, there was much disagreement with the assessment of question A1.

Table 7: Assessment of Decompositions

Question #	Max Points	Analyst 1	Analyst 2	Analyst 3	Analyst 4	mean
A1	5	4	0	3	3	3.0
A2	5	5	3	3	5	4.2
B1	6	6	6	4	5	5.4
B2	6	6	6	6	5	5.8

What we learned about the experiment

1. Thirty minutes appeared to be a time for re-grouping. In over half the experimental scenarios the analyst would stop interacting with the system and would review his/her notes and would re-read the scenario
2. Web-based interfaces have more features and require more training.
3. Using Camtasia to record the screens during the session allowed us to use this information during the debriefing session with the analysts. We also found this helpful during analysis if we wanted to recall specifics of a critical event. However, this type of analysis is extremely labor intensive.

Summary and Proposal for Next Experiment

In this experiment we were able to see improvement in the systems. We moved from primarily wizard based systems to systems that were able to handle a good portion of the queries alone or modified by a human.

The use of a common set of data eliminated some of the variability between systems. Using more realistic scenarios and giving the analysts more time to use the system allowed us to observe more in-depth. The web-based user interfaces have made us reconsider the definition of Dialogue. We submit that dialogue should be considered as:

Interaction with the system about content + interaction with the system to explore existing content (local interactions)

Interaction with the system about content consists of requests by the analyst, responses by the system and clarifications by either the system or the analyst. Requests for content can be explicitly made by the analyst; system facilitated requests (as in the “similar questions list”, or automated delivery by the system. Interactions concerning existing content would include visualizing results, reading documents supplied by the system,

Suggested metrics for the next experiment include:

- Q&A Dialogue Metrics
  - Efficiency
    - Average number of turns per successful thread
    - Number of non-clarification turns/total number of turns
  - Effectiveness
    - Relevance rating of answer for each question
    - Number of documents returned that were read or saved.
  - User Satisfaction metrics
    - Trust should be included
  - Ground truth
    - Percentage of documents retrieved from the CNS subsets specified by the CNS detailed decomposition.
- Local interactions
  - Usability
    - Errors/ misunderstandings in use
  - Utility
    - Time spent using features
    - Value added to process
  - User Satisfaction

Experiment Logistics

- Issues to be considered include:
  - Abolish the use of wizard and system modified responses
  - Ability to run multiple users at once
  - Move to all web-based user interfaces
  - Logging to be implemented by participating systems includes original logging of query/response + logging of use of local interactions (according to format developed by NIST)

As more systems move to web-based user interfaces, we will need more time for training. If we can support multiple users, we can train all analysts who will be using the system at the same session. Usability can be measured during this training session as well. This will give us a metric of “learnability” plus a usability metric during the actual use of the system.

Sessions should last at least one hour. We suggest that we try to expand these to two hours and that we do two scenarios per system. We could consider expanding use to three analysts for each system. We see no particular value in having analysts use more than one system.

We believe that we have learned a lot about how analysts interact with systems to accomplish an overarching task. The results of both Pilot studies have raised the bar for systems. As different teams decide to participate in this type of evaluation, they need to know the kinds of problems that have been encountered and enumerated in these prior studies so that they can make best use of that information in designing their systems.

Over the next few months we will work with researchers and the program management team to define the fall experiment for dialogue with the goal of defining a draft set of metrics for use.

Appendix A – Scenarios

Question set A: Training

What is the current status of India's Prithvi ballistic missile project?

Decomposition: Think about your strategy for working on this task and think about the information that you looked for and the information you found. Consult your notes if you like. Then, for each of the elements in the table below, check ‘Yes’ if your intended report would cover the material in the Element; otherwise, check ‘No’.

Yes	No	Element
		How should 'India' be identified? Pre-independence or post-independence, post-colonial or post-1947 India?
		What is 'Prithvi'? What does Prithvi mean? What class of missiles does Prithvi belong to? What is its range/payload, and other technical details?
		What is the meaning of 'status'? Does status mean research and development, flight-tests, user-trials, serial production, integration into the armed forces?

Question set A – Scenario 1

Despite having complete access, to this day UN inspections have been unable to find any biological weapons, or remnants thereof, in Iraq. Why has it proven so difficult to discover hard information about Iraq's biological weapons program and what are the implications of these difficulties for the international biological arms control regime?

Decomposition: Think about your strategy for working on this task and think about the information that you looked for and the information you found. Consult your notes if you like. Then, for each of the elements in the table below, check 'Yes' if your intended report would cover the material in the Element; otherwise, check 'No'.

Yes	No	Element
		Is there such a concept as "complete access" or are there inevitably limits to accessing sites and facilities? If there are such limits, can inspections in fact be carried out effectively; i.e., with an acceptable level of assurance that were there biological weapons and/or related systems, they would be found by inspectors?
		What is a biological weapon? Is it, for example, a quantity of pathogens or toxins, or is there more to it?
		What are the likely signatures of a national biological weapons program and how likely is it that inspectors from the outside would be able to detect them?
		What are the constituent parts of the "international arms control regime" in the context of biological weapons? Does it, for example, solely consist of the 1972 Biological and Toxin Weapons Convention (BWC), or is there more to it?
		Since Iraq was only a signatory (not ratifier) of the BWC during the time it was developing and producing biological weapons (1985-1991), were its actions in this regard contrary to international law? If not, did the international community have a different recourse to designate the Iraqi government as having violated international law or norms by having acquired biological weapons?

Question set A – Scenario 2

How have thefts impacted on the safety of Russia's nuclear navy, and has the theft problem been increased or reduced over time?

Decomposition: Think about your strategy for working on this task and think about the information that you looked for and the information you found. Consult your notes if you like. Then, for each of the elements in the table below, check 'Yes' if your intended report would cover the material in the Element; otherwise, check 'No'.

Yes	No	Element
		What sorts of items have been stolen? To what degree do different thefts put nuclear or radioactive materials at risk?
		What is meant by nuclear navy? Bases with active vessels carrying nuclear weapons only, bases with nuclear-powered vessels (and conventional weapons), bases with decommissioned nuclear-powered vessels, bases of civilian nuclear-powered vessels, or radioactive waste/spent nuclear fuel and/or fresh nuclear fuel sites?
		What does 'impact' mean? Is this only thefts of sensitive equipment, or does it include having an economic impact on the base (decreasing moneys for safeguards, etc.) or making it easier for officers to become embroiled in criminal pursuits (for instance, does the involvement of senior officers or organized crime groups have a greater impact than thefts by single recruits)?
		How does one define an increase or decrease in the problem? By dollar amounts, or degree of access to sensitive sites, or influence over men responsible for safeguarding sensitive sites?
		What specific instances of theft do we know about, and what are the sources of this information? What is the number of thefts that are likely to be reported (and how many are we probably missing)?

Question set B -- Training

Which countries contributed to Iraq's nuclear weapons program in the 1980s?

Decomposition: Think about your strategy for working on this task and think about the information that you looked for and the information you found. Consult your notes if you like. Then, for each of the elements in the table below, check 'Yes' if your intended report would cover the material in the Element; otherwise, check 'No'.

Yes	No	Element
		What are the countries that existed during the 1980s but changed status in the post-Cold War? E.g. West Germany, East Germany, Yugoslavia, Soviet Union. Did these countries contribute to Iraq's nuclear weapons program?
		What is meant by contribution? Is it assistance in technology, know-how, or direct transfers? Do we also mean ideological support to the development of the program?
		What is the relationship between Iraq and the contributor states?
		What are the landmark developments in such contribution and what are the sources of this information?

Question set B – Scenario 1

How does China's participation in nonproliferation regimes based on legally binding treaties (such as the NPT and the CWC) compare to its participation in informal multilateral agreements (such as the NSG and the MTCR)? What aspects of China's stated arms control policy would explain any difference?

Decomposition: Think about your strategy for working on this task and think about the information that you looked for and the information you found. Consult your notes if you like. Then, for each of the elements in the table below, check ‘Yes’ if your intended report would cover the material in the Element; otherwise, check ‘No’.

Yes	No	Element
		What is meant by 'participation'? Does this refer to involvement in negotiations, signing and ratifying treaties, active implementation of commitments or simple adherence to obligations?
		What is meant by 'nonproliferation regimes'? Does this only refer to multilateral arrangements or are bilateral agreements or discussions included in the general definition?
		What is the difference between formal treaties and informal agreements? Is there a difference in the level of responsibility for States participating in a treaty versus an informal agreement? Does China's participation correlate with the level of responsibility required?
		To what treaties and agreements is the government of China party? What led China to agree to these treaties? What do Chinese public statements demonstrate about China's support or opposition to certain nonproliferation agreements?
		What is meant by 'Chinese arms control policy'? Does this only include official Chinese statements, or are reports in leading Chinese newspapers, like People's Daily, China Daily or PLA Daily also included?
		What agreements or treaties has China not joined? Are there agreements that China adheres to but has not officially joined? Are there types of agreements or aspects of agreements that the Chinese government consistently does not support?

Question set B – Scenario 2

What kind of assistance has North Korea received from the USSR/Russia for its missile program?

Decomposition: Think about your strategy for working on this task and think about the information that you looked for and the information you found. Consult your notes if you like. Then, for each of the elements in the table below, check ‘Yes’ if your intended report would cover the material in the Element; otherwise, check ‘No’.

Yes	No	Element
		What is the USSR/Russia? Is this the Soviet/Russian government? Does it include private firms, state-owned firms, educational institutions, and individuals?
		What is North Korea? Is this the North Korean government only? Does it include private firms, state-owned firms, educational institutions, and individuals?
		What is assistance? Is it the transfer of complete missile systems, licensing agreements, components, materials, or plans? Is it the training of personnel? What kind of training? Does transfer include data, and, if so, what kind of data? Does transfer include financial assistance, and, if so, what kind of financial assistance?
		What are the missiles in the North Korean inventory? Are any based upon Soviet/Russian designs? If so, which ones? What was the development timeline of the missiles? Did any timeline differ significantly from others? Did North Korea receive assistance from other sources besides USSR/Russia to develop these missiles?
		When did North Korea receive assistance from the USSR/Russia? Was any intended assistance halted, stopped or intercepted?
		What are the sources of information? Are the sources reliable? Is some information contradictory?

Appendix B – Training Materials

Overall Study Description

AQUAINT Dialogue Evaluation

The **purpose** of this experiment is to investigate information systems that engage the user in a dialogue that hopefully results in the retrieval of useful, relevant information.

You will be given information seeking scenarios.

- **Your role** – intelligence analyst.
- **Your tasking** – prepare a report on the topic for an upper-level client.
- **Your client/customer** – upper-level management, e.g., Deputy Secretary
- **Task duration** – 1 week to gather information, analyze it and synthesize it into a report.
- **Today's work:** Begin the process of gathering pertinent information in as many of the sub-areas as you can. You are free to break up the problem as you see fit. Often a useful strategy is to verify terminology used in the tasking.

You will have one hour per scenario to interact with the system to locate information. There are three stopping conditions:

1. You are satisfied with the information the system has delivered and feel ready to write your report..
2. You are not satisfied with the information but you don't feel that any more progress can be made.
3. The 1 hour is up. We will notify you when this occurs.

The user interface is similar to a 'Chat' window. You will be able to enter you questions or statements about your information needs and the system will reply. A separate packet of directions for using the interface is included.

As you collect information, you will want to keep track of it. You may use paper-and-pencil or a software application, e.g., Word, Notepad and WordPad. Circle your preference from the following list.

1. paper and pencil
2. Word
3. Notepad
4. WordPad

Some one will be in the room to help you with any problems you have. Please let us know when you would like a break. If possible, breaks should be taken between scenarios.

There is a training scenario, labeled T. We will work through this scenario until you feel comfortable using the user interface. The experiment scenarios are numbered E1 and E2. Please complete them in this order. After you have worked with each scenario, you will be asked to fill out a questionnaire rating various aspects of the system.

Instructions for using the system interface were tailored for each system. Once the subjects had read the above tasking information, the proceeded to read the interface instructions.

Appendix C – Questionnaire

AQUAINT Dialogue Evaluation Questionnaire

Part 1: Please indicate your impressions of the system based on the following 7-point scales:

1	2	3	4	5	6	7
Pleasing						Irritating
Comment						

1	2	3	4	5	6	7
Helpful						Hindrance
Comment						

1	2	3	4	5	6	7
Understood my questions						Did not understand my questions
Comment						

1	2	3	4	5	6	7
Easy to understand						Incomprehensible
Comment						

1	2	3	4	5	6	7
Would use regularly						Would never use
Comment						

1	2	3	4	5	6	7
Handled context well						No understanding of context
Comment						

1	2	3	4	5	6	7
Very natural dialogue						Unnatural, difficult dialogue
Comment						

1	2	3	4	5	6	7
Easy to understand what to say						Difficult to understand what to say
Comment						

1	2	3	4	5	6	7
Natural turn taking						Difficult to know when it was my turn
Comment						

1	2	3	4	5	6	7
Verbose						Terse
Comment						

1	2	3	4	5	6	7
System was in control						I was in control
Comment						

Part 2:

Prior to this task, how much did you know about the task domain?

1	2	3	4	5	6	7
Expert						Novice
Comment						

How much of your prior knowledge did you apply to the solution of this task?

1	2	3	4	5	6	7
Little						Much
Comment						

How confident are you in the accuracy of your assessment of the task problem?

1	2	3	4	5	6	7
Not at all confident						Very confident
Comment						

How confident are you that you have covered the important components of this problem/task?

1	2	3	4	5	6	7
Not at all confident						Very confident
Comment						

Would you use this system to complete your report, given your experience so far?

1	2	3	4	5	6	7
Would NOT use						Would definitely use
Comment **REQUIRED]**						

Appendix D – Initial Queries

	Scenario			
Subject	**A1**	**A2**	**B1**	**B2**
A	What evidence has the UN found on Iraq's biological weapons?	How have thefts impacted on the safety of Russia's navy?	China's participation in nonproliferation regimes based on legally binding treaties (such as NPT and the CWC) compare to china's participation informal multilateral agreements (such as NSG and the MTCR)	What kind of assistance has North Korea received from the USR/Russia for its missle program
B	imagery of mobile Iraqi chemical labs	In recent decades since the fall of the communist party in Russia what has been the theft rate in regards to stolen nuclear secrets	What is meant by the term nonproliferation.	What type of military assistance has North Korea received from the USSR before it became Russia in the modern nuclear age.
C	What items have been stolen from the Russian Navy	What evidence exists concerning Iraq's biological weapons program	What is China's publicly stated policy on proliferation?	What ballistic missiles does North Korea possess?
D	Looking for information on UN weapons inspections in Iraq -- when and how the inspections have been carried out?	Looking for information on the occurence of thefts of Russian navy equipment.	Looking for information concerning China's participation regarding the NPT?	How has the USSR assisted North Korea with its missile program?

Appendix E – All Queries

Scenario A1

Analyst 1

- Looking for information on UN weapons inspections in Iraq -- when and how the inspections have been carried out?
- Which sites have been inspected?
- How have the inspections been conducted? Who chooses the site? How much notice give?
- Are the inspectors accompanied by Iraqi personnel? Are their movements throug the country monitore?
- Are the inspectors able to check any suspected storage sites that are military, such as palaces, schools?
- How was the list of 100 sites compiled?
- When UN inspectors arrive "unannounced", do the personnel at the site seem to be surprised or prepared?
- Have Iraqi officials hindered the movements of the inspectors in any way?
- Any other difficulties reported by the inspectors?
- Has the UN been able to use U-2 flights for its inspections at all?
- How many U-2 flights have been flown?
- What evidence of biological weapons and/or weapons development has been found?
- Besides physical inspections, air, water and vegitation sampling, and possible U-2 flights, what other methods has the UN used to detect biological weapons?
- Have they actually conducted interviews with Iraqi personnel?
- Is there any evidence that Iraqi personnel who have been interviewed have had any problems following any interview?
- Retribution, interrogation, ill treatment for having spoken with inspectors.
- Have the inspectors encountered any difficulties in travelling from one site to another?

Analyst 2

- Imagery of mobile Iraqi chemical labs
- Known Iraqi scientist working in military facilities
- Top Iraqi officals who may have knowledge of chemical and biological weapons development in Iraq
- Known suppliers of chemical and biological casings/materials sent to Iraq within the past 10 years
- In recent years since the UN inspectors have returned to Iraq to investicate the possible development of chemical/biological weapons have there been increased military flights to Syria
- Of the top Iraqi officials and scientist how many of their family and friends have gone missing or suffered severe medical disablities
- What are the known underground military facilities that are in Iraq and have they been visited by the UN inspectors
- Since Korea has admit that they have nuclear weapons and it is known that the development of these weapons were developed at underground facilities can we find parallels to the Iraqi's biological weapons development in underground facilities
- Since the UN inspector have gone to Iraq have they received free access to the Iraqi facilities as known chemical/biological producers reported to the UN by various intelligence communities
- What can the UN do to protect top Iraqi officals and scientist who want to testify that there is a chemical/biological weapons being developed in Iraq or other axis of evil countries (Iran, Korea, Iraq and now Syria)
- How can the UN deter other countries from supplying chemical/biological weapons to other countries with the intent to develop chemical/biological weapons
- What kind of sanctions has the UN put on other countries that violate the arms control agreement to develop weapons of mass destruction
- What kind of message has the UN sent with the international community with Iraq and Korea on there development of weapons of mass destruction.

- Based on the Iraqi and Korean development of weapons of mass destruction what steps have the UN chemical/biological inspections taken to update or restructor their inpections program.
- Since Operations Iraqi freedom and the assured victory of the coalition forces will it be easier for Iraqi scientist to come forward with the down fall of the Iraqi regime
- Recent news from Operations Iraqi freedom has indicated that the biological weapons have been sent to Syria or buried with the capture of top Iraqi officials will the locations of the weapons stores be forthcoming.
- Operations Irawi freedom has indicated that Syria may now possess some of Iraq's biological weapons what will the UN do to Syria. Will the UN request that inspections take place in Syria or will the intelligence community obtain information from other sources

Analyst 3

- What items have been stolen from the Russian Navy
- Has Russia's nuclear navy experienced a rise in accidents between 1992-2002 compared with the previous decade
- What has the trend in mishaps been in Russia's nuclear navy over the period 1992-2002

Analyst 4

- What evidence has the UN found on Iraq's biological weapons?
- Why hasen't the UN found biological weapons?
- What has hindered the UN in it's search for biological weapons?
- what facilities has Iraq used to produce biological weapons?
- What is the evidence the US has that Iraq has concealed weapons?
- What indications does the US have that Iraq has moved it's biological weapons to another country?
- Where have biological outbreaks been reported?
- What has been Iraq's reaction to UN inspectors?
- Do the UN inspectors have the proper equipment to detect biological weapons?

Scenario A2

Analyst 1

- Looking for information on the occurence of thefts of Russian navy equipment.
- How often do thefts occur?
- What are radionuclides and are they hazardous?
- How are radionuclides used in nuclear weapons?
- Why are these substances kept by the navy? How are they stored?
- How has the navy used them for its submarines?
- So, are radionuclides fuel or waste?
- Have there been thefts of any other type of equipment or supplies from the Russian navy?
- How often have nuclear submarines been disabled as the result of thefts?
- How have the thefts impacted on the overall supply of fuel for nuclear submarines?
- Is there any information on how the thefts have impacted on the supply of fuel for nuclear subs in Russia?
- Have any subs other than the one prviously mentioned been disabled by theft?
- Have any thefts resulted in dangerous outcomes?
- What is the number of thefts from the Russian nuclear navy during the past year?
- What was the rate of theft the year before?
- How many of the thefts were from the navy?

Analyst 2

- In recent decades since the fall of the communist party in Russia what has been the theft rate in regards to stolen nuclear secrets
- How has the Russian government tried to deter stolen or selling of nuclear secrets
- Who has been the primary suspect in most of the Russian stolen nuclear secrets
- Have any of the stolen Russian nuclear secrets been sold to terrorist organizations or countries seeking to expand their nuclear capabilites
- What has been the punishment by the Russian government on individuals or groups caught selling stolen Russian nuclear secrets
- How has the Russian Navy suffered with nuclear secrets stolen
- Due to the many mishaps over several decades with the Russian nuclear submarine how has the Russian Navy taken preventive measures secure their nuclear secrets
- Is theft the primary cause of the Russian nuclear navy mishaps or is it lack of safety practises by the Russian navy
- How often do the Russian Navy practise safety procedures on a nuclear vessel
- How often do Russian scientist or top Russian officials defect to other countries
- Are the defectors tracked and what is the Russian government doing to prevent the defectors from giving up the countries nuclear secrets
- Since the fall of the communist party in Russian during which Russian premier was there increase in stolen nuclear secrets and during which premier's regime did it reduce
- What did the Russian government do when the increase in stolen navy nuclear secrets was on the rise
- Has the Russian Nuclear navy theft decrease and what were the factors for the decline
- What has been the safety record of the Russian Nuclear Navy since it's development.
- What organization within the Russian Nuclear Navy oversees safety issues and who is accountable for the safeguarding of nuclear secrets
- What type of theft has been reported by the Russian nuclear navy and how has that impacted their safety
- How does the theft of the Russian nuclear navy impact the rest of the world more specifically the United States
- Who has been the known buyers of stolen Russian navy nuclear secrets and what price were they willing to pay
- Has any country/organizations/individuals been brokers for other countries/organizations/individuals seeking stolen Russian nuclear navy secrets
- Have there been any whistle blowers within the Russian Navy of countries/organizations/individuals seeking to recruit or buy information about the Russian Navy nuclear programs
- Has the Russian government admit that this is a increasing problem within the Russian Navy of stolen nuclear secrets or has the problem declined where it is no longer a factor
- How would the Russian Nuclear Navy classify theft to their program
- How would the Russian Navy classify a safety problem to their navy, is it loss of life, loss of equipment, loss of classified material or environmental conditions
- What statistics do the Russian Navy have that are comparitive to the increase or decrease in theft measured against the Russian Navy safety record

Analyst 3

- What evidence exists concerning Iraq's biological weapons program
- What is needed for a country to have a biological weapons program

Analyst 4

- How have thefts impacted on the safety of Russia's navy?
- Has the theft of equipment from the Russian navy increased or r3educed over time.?

- Have there been any accidents reported as a result of thefts from the Russian Navy?
- How has the effectiveness of the Russian Navy been impacted due to the theafts?
- Have any Russian ships been taken out of service due to thefts?
- What equipment has been stolden from the Russian navy?
- Have any fatal accidents been reported for the Russian navy.
- Has the Russian navy reported any fatal accidents?
- Safety training for Russian navy.
- What is the United States involvement in the Russian navy safety program?
- What countries have received stolen material from the Russian navy?

Scenario B2

Analyst 1

- How has the USSR assisted North Korea with its missile program?
- Prior to 1990, how has the Soviet Union, or USSR, assisted North Korea with its missile program?
- Since 1990, how has the USSR or Russia assisted North Korea's missile program?
- Information on missile technology assistance given to North Korea from USSR or Russia?
- Any evidence of transfer of technological information about missiles from USSR or Russian to North Korea?

Analyst 2

- What is meant by the term nonproliferation.
- What does the acronym NPT, CWC, NSG and MTCR mean.
- Define China's arms control policy.
- Is this not a contridiction for China since they are the main suppliers or North Korea's nuclear weapons program.
- Who are the members/signers of the NPT, CWC, NSG, and MTCR.
- The members of the NPT, CWC, NSG, and MTCR which ones have nuclear, biological, and chemical weapons.
- Define NPT, CWC, NSG, and MTCR. What are the consequences of breaking these treaties/agreements.
- How do you compare China's participation among the different treaties, NPT, CWC, NSG and MTCR.
- Since China does not belong to NSG has it exported nuclear and nuclear-related technologies to other countries and if so does this not go against their arms control policies.
- Then why sign the NPT, CWC and MTCR if China's Leaders wants complete prohibition and thorough destruction of all weapons of mass destruction--nuclear, biological, and chemical.
- So China wanted to go along with the rest of the world with no intent of reducing its nuclear weapons or exporting to other countries who have the currency to buy such technologies.

Analyst 3

- What is China's publicly stated policy on proliferation?
- How has China complied with the Nonproliferation Treaty?
- Has China complied with the Chemical Weapons Convention?
- How has China complied with the MTCR?
- What is China's official position toward the NSG?
- Who is Sha Zukang?
- Find all references to Sha Zukang.

Analyst 4

- China's participation in nonproliferation regimes based on legally binding treaties (such as NPT and the CWC) compare to china's participation informal multilateral agreements (such as NSG and the MTCR)

- China's participation in nonproliferation regimes based on legally binding treaties (such as NPT and the CWC) compare to china's participation informal multilateral agreements (such as NSG and the MTCR)
- NPT membership

Scenario B2

Analyst 1

- Looking for information concerning China's participation regarding the NPT?
- What is CWC as relating to nonproliferation regimes?
- Information on China's participation in NSG?
- How about China and MTCR?
- What is IAEA?
- What is China's stated policy on arms control?

Analyst 2

- What type of military assistance has North Korea received from the USSR before it became Russia in the modern nuclear age.
- The Russians supplied missile technology/knowledge transfer and did not supply the North Koreans with equipment or materials for missile development.
- Did the Russians supply the North Koreans with equipment or nuclear materials in development of their missile program.
- How long and how recent has this been going on.
- Does the Russian belong to any nonproliferation treaties banning the transfer and exportation of nuclear material or knowledge.
- What consequences has the Russian received from the NPT, CWC, NSG and MTCR for violating these agreements.
- What kind of sanctions were imposed on Russia and have they been lifted.
- What kind of compensation did the Russians receive from North Koreans for assistance with their missile program.
- The latest news reports that N. Korea can barely feed its people or provide electricity. Strong sanctions from the world has severely hurt its economy where do the North Koreans get their finances from.
- Do the North Koreans get any financial support from any other countries, e.g. China.
- Define North Korea's missile program.
- Then why would the Japanese give financial support to North Korea if they were in essence in targeting range from the North Korean missiles - Is it for appeasement.
- Other than financial compensation what else has the Russians received from the North Koreans in helping develop their missile program. Will the Russians receive access to that country to gather information about countries in that region.

Analyst 3

- What ballistic missiles does North Korea possess?
- Have Russian missile experts participated in Nodong development?
- What specific assistance did Russian missile experts provide North Korea?
- What is the Taepodong?
- What is the range of the Taepodong-2?
- Have Russian experts helped North Korea develop the Taepodong?
- In what aspects of missile design does the Makayev Institute engage?
- What are the technical specifications of the Russian SS-4 missile?

Analyst 4

- What kind of assistance has North Korea received from the USR/Russia for its missle program
- Russian monetary assistance to North Korea for missile program